<Bioinformatics Center>Mathematical Bioinformatics

Bioinformatics Center
– Mathematical Bioinformatics –

http://www.bic.kyoto-u.ac.jp/takutsu/

Prof
AKUTSU, Tatsuya
(D Eng)

Assist Prof
HAYASHIDA, Morihiro
(D Inf)

Assist Prof
TAMURA, Takeyuki
(D Inf)

Program-Specific Res
KOYANO, Hitoshi
(D Agr)

Students

KAMADA, Mayumi (D3)
NAKAJIMA, Natsu (D3)
NARITA, Yuki (D3)
ZHAO, Yang (D2)
LU, Wei (D2)
HASEGAWA, Takanori (D1)
MORI, Tomoya (D1)
RUAN, Peiying (D1)
UECHI, Risa (D1)
SHI, Dongyue (M2)
JIRA, Jindalertudomdee (M2)
CONG, Xiao (M1)
HWANG, Jaewook (M1)

Visitor

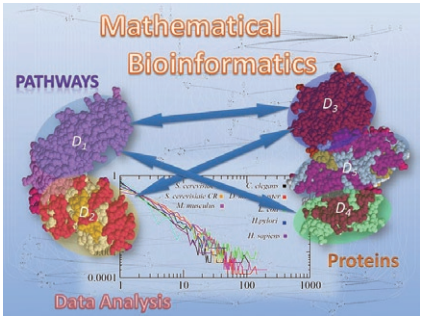
Ms JIANG, Hao The University of Hong Kong, China, P.R., 4 June–15 August

Scope of Research

Due to rapid progress of the genome projects, whole genome sequences of organisms ranging from bacteria to human have become available. In order to understand the meaning behind the genetic code, we have been developing algorithms and software tools for analyzing biological data based on advanced information technologies such as theory of algorithms, artificial intelligence, and machine learning. We are recently studying the following topics: systems biology, scale-free networks, protein structure prediction, inference of biological networks, chemo-informatics, discrete and stochastic methods for bioinformatics.

KEYWORDS

Scale-free Networks
Boolean Networks
Grammar-based Compression
RNA Secondary Structures
Chemical Graphs



Selected Publications

Akutsu, T.; Kosub, S.; Melkman, A. A.; Tamura, T., Finding a Periodic Attractor of a Boolean Network, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9,** 1410-1421 (2012).

Akutsu, T.; Fukagawa, D.; Jansson, J.; Sadakane, K., Inferring a Graph from Path Frequency, *Discrete Applied Mathematics*, **160,** 1416-1428 (2012).

Nacher, J. C.; Akutsu, T., Dominating Scale-free Networks with Variable Scaling Exponent: Heterogeneous Networks Are Not Difficult to Control, *New Journal of Physics*, **14**, 073005 (2012).

Hayashida, M.; Ruan, P.; Akutsu, T., A Quadsection Algorithm for Grammar-Based Image Compression, *Integrated Computer-Aided Engineering*, **19,** 23-38 (2012).

Kato, Y.; Sato, K.; Asai, K.; Akutsu, T., Rtips: Fast and Accurate Tools for RNA 2D Structure Prediction Using Integer Programming, *Nucleic Acids Research*, **40,** W29-W34 (2012).

Network Completion Using Dynamic Programming and Least-Squares Fitting

Analysis of biological networks is one of the central research topics in bioinformatics and computational systems biology. Recently, we proposed a concept of "Network Completion", which is to make the minimum amount of modifications to a given network so that the resulting network is most consistent with observed data. In our previous work, we studied the computational complexity of network completion under Boolean models.

In this work, we propose a novel method, DPLSQ, for completing genetic networks using gene expression time series data. Different from our previous work, we employ a model based on differential equations and assume that expression values of all nodes can be observed. DPLSQ is a combination of least-squares fitting and dynamic programming and one of its important features is that it can output an optimal solution (i.e., minimum squared sum) in polynomial time if the maximum indegree (i.e., the maximum number of input genes to a gene) is bounded by a constant. Although DPLSQ does not automatically find the minimum modification, it can be found by examining varying numbers of added/deleted edges, where the total number of such combinations is polynomially bounded. If a null network (i.e., a network having no edges) is given as an initial network, DPLSQ can work as an inference method for genetic networks. In order to examine the effectiveness of DPLSQ, computational experiments were conducted using both artificially generated and real gene expression time series data.

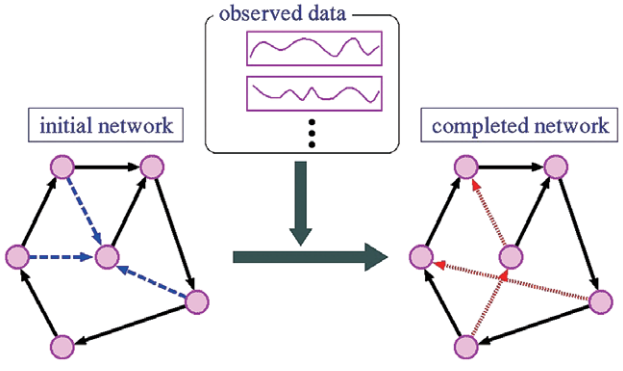


Figure 1. Network completion by addition and deletion of edges. Dashed edges and dotted edges denote deleted edges and added edges, respectively.

Predicting Protein-RNA Residue-base Contacts Using Two-dimensional Conditional Random Field

To dissect the interactions between proteins and RNAs leads novel findings of molecular networks and functions in cellular systems. In terms of the interactions between amino acid residues in proteins, it is generally accepted that residues at interacting sites have coevolved with the corresponding residues in the partner protein to keep the interactions between proteins. In our previous work, based on this hypothesis, we calculated mutual information (MI) between residues from multiple sequence alignments of homologous proteins to identify residue-residue contact pairs in interacting proteins, and combined it with a discriminative random field (DRF) approach, which is a special type of conditional random fields (CRFs). Recently, the evolutionary correlation of the interactions between residues and DNA bases has also been found in certain transcription factors and the DNA-binding sites.

In this work, we employ CRFs to predict the interactions between protein residues and RNA bases. Furthermore, we introduce labels of amino acids and bases as features of a simple two-dimensional CRF instead of DRF. In addition, we examine the utility of L1-norm regularization (lasso) for CRF. To evaluate our method, we perform computational experiments of several interactions between the Pfam domains and Rfam entries, and calculate the average AUC (Area under ROC Curve) score. The results suggest that our CRF-based method using MI and labels with lasso is useful for further improving the performance, especially provided that the features of CRF are successfully reduced by the lasso approach.

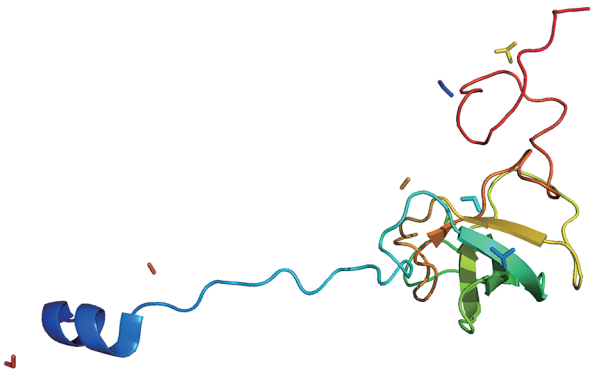


Figure 2. Protein RS12_THET8 of PDB code '1yl4' and the atoms of RNA M26923 within 3 Å of the protein.